Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications

Oussama Ahmia, Nicolas Béchet, Pierre-François Marteau

Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications

Oussama Ahmia[1,2], **Nicolas Béchet**[1], **Pierre-François Marteau**[1]
[1]IRISA, Université Bretagne Sud
Campus de Tohannic, 56000 Vannes FRANCE
[2] Jurismarchés, 2 Place Saint-Pierre, Nantes, 44000, FRANCE
{firstname.name}@irisa.fr

Abstract

The European "Tenders Electronic Daily" (TED) is a large source of semi-structured and multilingual data that is very valuable to the Natural Language Processing community. This data sets can effectively be used to address complex machine translation, multilingual terminology extraction, text-mining, or to benchmark information retrieval systems. Despite of the services offered by the user-friendliness of the web site that is made available to the public to access the publishing of the EU call for tenders, collecting and managing such kind of data is a great burden and consumes a lot of time and computing resources. This could explain why such a resource is not very (if any) exploited today by computer scientists or engineers in NLP. The aim of this paper is to describe two documented and easy-to-use multilingual corpora (one of them is a parallel corpus), extracted from the TED web source that we will release for the benefit of the NLP community.

Keywords: Multilingual corpora, Parallel Corpus, Call for Tender, European Languages, Natural Language Resource

1. Introduction

The world procurement generates daily large amounts of data, that represent useful knowledge for business intelligence tasks.
Among other sources, the European "Tenders Electronic Daily" (TED) system, publishes approximately 1,700 tenders five times a week in the TED servers[1]. More precisely, the publication office of the TED publishes about 460,000 contract notices per year, among which 175,000 tenders worth around 420 billion euros. These raw data are available as bulk downloads of these data that contain contract notices for tender data in XML format with a complex structure. Unfortunately, different versions of the XML data structure from year to year have been used, making the aggregation of the different bulks of data difficult. Furthermore, the collected documents are associated to a variable number of translations as well as variable sets of meta data that is used for indexing. Consequently, the collecting and managing of such data is a great burden and consumes a lot of time and computing resources.

The aim of this paper is to describe a processed version of this database in a raw text format that can be directly and easily used for text mining and natural language processing tasks. We also aim at making this processed dataset available to the scientific community and can be downloaded[2] along with a simple Python API for easier manipulation.

The provided dataset is declined into two sub-datasets created from the TED's documents that have been published between January 2011 and August 2017.

1. The first sub-dataset, fd-TED, is a (multilingual) corpus or aligned translated documents. It contains around 3 millions of documents translated to 24 languages (DA, DE, EN, ES, FI, FR, EL, IT, NL, PT, SV, CS, ET, HU, LT, LV, MT, PL, SK, SL, GA, BG, RO, HR). This dataset can be used as a benchmark for supervised classification or for training machine learning models applied to business intelligence application.

2. The second sub-dataset, par-TED, consists of the aligned sentences of translated texts extracted from the fd-TED corpus. It can be used for machine assisted translation of juridical and technical documents, or the extraction of multilingual terminology. This corpus is composed with 4 millions of unique sentences translated to at least 23 languages.

The two sub-datasets, fd-TED and par-TED, will be updated in a regular basis to keep tracks of the new calls for tender published by the EU states.

We also provide an API, to download the new updates and to support an easy access to the data. This is done through the use of filters that can be applied on the meta data, basically the language(s), the hierarchical level(s) of the Common Procurement Vocabulary (CPV) codes, the type of processed texts, and so on.

2. The (full-document) fd-TED corpus

The fd-TED corpus is built from the full content of the documents extracted from the TED platform. Each document of the corpus belongs to a hierarchy that is succinctly described below.

[1]http://ted.europa.eu/TED/main/HomePage.do
[2]`https://github.com/oussamaahmia/TED-dataset`

Level in the hierarchy	1	2	3	4	5	6	7	8	9
Count	1,868,420	433,111	231,167	144,393	115,487	45,656	30,792	21,694	16,727
Cumulative	2,907,447	1,039,027	605,916	374,749	230,356	114,869	69,213	38,421	16,727

Table 1: Number of documents for each level of the CPV code

Language	DE-ES	DE-IT	EN-DE	EN-ES	EN-FR	EN-IT	FR-DE	FR-ES	FR-IT	IT-ES
Count	425,797	428,097	425,893	425,808	426,027	425,856	429,039	425,797	425,803	425,797

Table 2: Number of documents fully translated for some pairs of languages.

2.1. Common Procurement Vocabulary

Common Procurement Vocabulary (CPV) [3] is the "thesaurus" that defines the subject matter of public contracts, allowing companies to easily find public procurement notices according to their areas of expertise. The main CPV vocabulary is based on a hierarchical structure (a tree structure) comprising codes of up to 9 digits (the ninth digit serves to check the previous digits). The CPV code consists of 8 digits that encodes 5 hierarchical subdivisions as follows:

1. The first two digits identify the divisions (XX000000-Y), e.g. "industrial machinery".
2. The first three digits identify the groups (XXX00000-Y), e.g. "Machine tools".
3. The first four digits identify the classes (XXXX0000-Y), e.g. Metal-working machine tools.
4. The first five digits identify the categories (XXXXX000-Y), e.g. "Hydraulic presses".
5. Each of the last three digits gives a greater degree of precision within each category.

For Example:
42000000 is the code for "industrial machinery", 42600000 is for "Machine tools", 42630000 for "Metal-working machine tools" and 4263600 is for "Hydraulic presses".
Table 1 presents the number of documents for each level of the CPV codes by taking into account the last hierarchical level (the 8 digits of the CPV code)

2.2. The documents

The documents are published in 24 languages of the EU. They can be fully translated to the 24 languages (Table 2 and 3.) or partially translated (in most of the cases the object of the document and the lots [4] are translated).
The dataset that we provide is presented as a multilingual corpus that can be exploited for supervised hierarchical classification or Cross-Language Text Classification (Olsson et al., 2005).
The XML schema comes in different versions (R2.0.9 and R2.0.8), hence the needed fields are extracted using the parser corresponding to each version. Then the CPV codes are corrected if additional characters are found. The descriptions fields (named "desc") are created from the aggregation of several XML elements that are checked for administrative sentences using a classifier trained with manually tagged dataset by Jurismarches[5] experts (the accuracy of the classifier is 98%). The raw text is created for each available language by converting the XML into text records. In case of any error in this step, the text is downloaded directly from the TED's website or converted using TED's online API. The filtered text is created by ignoring all the XML entities dealing with administrative information (some XML elements will always contain only administrative content) and filtering the mixed elements using the classifier to get rid of the administrative content.
Knowing that the procurement notices contain legal and administrative information that are not fundamental for understanding the core business of the consultation, as filtered sentences tends to introduce a lot of noise if the interest is upon valuable business informations present in call for tenders (conditions relating to the contract, deposits and guarantees required...).
With the help of experts in public markets (Jurismarches, we provide a filtered version of each document that only contains the description of the supplies.
Example of core business information:

- Installation of doors and windows and related components.

Example of legal and administrative information that has been filtered out:

- Candidates (all partners in the case of a consortium) shall prove that they have the legal capacity to perform the contract by providing (...)

The data structure of the documents contained in the fd-TED corpus is presented in Figure 1.

2.3. Classification Example

As an example of supervised classification, Table 3 shows the results of a classification using Linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) and a bag of words representation. From a random sub-sample of 200K English and French documents extracted from the fd-TED corpus, we randomly split our data into into 75%

[3] COMMISSION REGULATION (EC) No 213/2008 of 28 November 2007

[4] Tenders are generally advertised with a global title, the object, and some of them are divided into lots, each having it's own title

[5] https://www.jurismarches.com/

```
{"ref":0000-0000 #The document ID in the TED database.
 "origin_ln":"" #The original language of the document.
 "list_ln": [] #the list of languages in which the document is translated.

 "document":{
     "EN":
         {"title":"Document Title" #The title of the document.
          "CPV":['00000000'] #The list of CPVs codes of the document
          "desc": "description of the project"
                       #additional information about the project.
          "lots": [ #list of the parts of the project.
                    "title":"Title of the lot"
                    "CPV": ['00000000'] #the CPVs codes of the lot.
                    "desc": "description of the lot"
                         #additional information about the lot.
          ]
          "raw": "the raw text of the document" # full text
          "filtered": "the processed document" # the text without the
          administrative information.

         }
 }}
```

Figure 1: Format of the documents for the processed dataset fd-TED.

for training and 25% for testing. We have used for this experiment the first hierarchical levels of the CPV codes, namely the two first digits.

The combination of the model trained on the English version of the documents and the French one using a max rule (Kittler et al., 1998) increases significantly the accuracy of this classification task.

Language	Accuracy
FR	59%
EN	65%
EN+FR	68%

Table 3: SVM classification results

3. The (parallel) par-TED corpus

Alongside with the fd-TED corpus, we provide a multilingual aligned corpus in the form of a set of parallel sentences with at least 1.2 million unique sentences translated to at least 23 languages. This corpus is created by aligning the XML trees for each language. Some XML elements are ignored (such as Phone numbers, email, addresses, etc). Then the repeated sentences are deleted.
Below is an example of aligned sentences for the EN,FR,ES,and IT languages.

- FR: Travaux de finition et de rénovation pour le complexe tokamak, le bâtiment d'assemblage et tous les bâtiments voisins.
- EN: Finishing and retrofit works for the Tokamak complex, assembly hall and all surrounding buildings.
- ES: Obras de modernización y finalización del complejo y taller de montaje del Tokamak y de los edificios colindantes.
- IT: Lavori di rifinitura e di ammodernamento per il complesso Tokamak, il reparto di assemblaggio e tutti gli edifici circostanti.

The data structure for the par-TED corpus is presented in Figure 2.
As an example, we have built word embeddings for the EN and FR languages to show the potentiality of this corpus in a multilingual terminology extraction application.
From Table 4 and Table 5 we can see that using a cosine similarity on Word2Vec representations (Mikolov et al., 2013) built on this corpus, we get comparable results regarding the word similarity on excerpts of common and proper nouns for the two tested languages.

4. Conclusion

The "Tenders Electronic Daily" (TED) is a large source of semi-structured and multilingual document widely underused by the NLP community, mainly due to the burden and costs associated to the collecting and formatting of the data. To ease the exploitation of this resource either for text mining or machine translation tasks, we have presented a packaging of these data (along with a Python API to access it) that can be freely downloaded and used by scientists and engineers to benchmark or solve some of their NLP problems. Few toy application examples have been also detailed to highlight the usefulness of this this resource and the type of services it can provide.

```
{"ref":0000-0000 #The document ID in the TED database.
 "origin_ln':"" #The language of the source document.
 "sent_id": #Sentence id in the document.

 "sentences":{ #List of the translations.
     "EN":"...",
     "FR":"...",
     "ES":"...",
     ...
          }
 }}
```

Figure 2: Format of the documents in the par-TED corpus.

Word	Similar Words	Similarity	Word	Similar Words	Similarity	Word	Similar Words	Similarity
Linux	citrix	0.81	Twitter	facebook	0.85	Lawyer + Advice	legal	0.70
	unix	0.81		social media	0.82		matters	0.69
	server	0.80		blogs	0.78		legal matters	0.68
	vmware	0.80		web chat	0.69		advisers	0.66
	microsoft	0.78		press releases	0.68		disputes	0.66
	windows server	0.77		youtube	0.67		matters arising	0.65
	weblogic	0.77		newsletter	0.66		legal advice	0.65
	oracle	0.77		text messaging	0.65		specific issues	0.65
	ms sql	0.77		direct mail	0.65		lawyers	0.65
	red hat	0.77		google	0.65		advice guidance	0.65

Table 4: Example of some most similar words using Word2Vec embedding and cosine similarity on English corpus

Word	Similar Words	Similarity	Word	Similar Words	Similarity	Word	Similar Words	Similarity
Linux	windows	0.85	Twitter	facebook	0.90	avocat + conseil	representation	0.75
	redhat	0.83		instagram	0.86		conseils	0.74
	unix	0.83		netflix	0.84		droit social	0.74
	mac os	0.82		snapchat	0.82		assistance juridique	0.73
	citrix	0.81		google	0.81		avocats	0.73
	serveurs	0.80		tweets	0.80		conseils juridique	0.72
	microsoft	0.79		youtube	0.80		representation juridique	0.72
	ibm	0.79		linkedin	0.77		contentieux	0.72
	windows server	0.79		maddyness	0.77		representation devant	0.71
	env windows	0.79		tweet	0.77		conseil juridique	0.71

Table 5: Example of some most similar words using Word2Vec embedding and cosine similarity on French corpus

SL	SK	DE	EL
451.1K/1.2M/433.0K	452.0K/1.4M/491.6K	849.4K/6.4M/1.5M	461.5K/1.7M/526.7K
LT	GA	PT	FI
457.3K/1.6M/496.3K	425.8K/868.6K/359.9K	450.7K/1.0M/371.1K	472.2K/1.3M/696.6K
MT	SV	LV	HU
425.8K/913.8K/352.1K	499.8K/1.3M/592.8K	443.3K/1.1M/420.9K	457.4K/2.5M/691.8K
EN	DA	FR	ES
674.8K/4.2M/824.9K	461.3K/1.4M/545.6K	1.1M/11.4M/1.2M	560.3K/2.1M/510.0K
RO	PL	HR	IT
483.1K/3.5M/567.8K	739.8K/11.2M/988.6K	288.3K/765.3K/314.9K	544.9K/2.9M/677.3K
NL	CS	ET	BG
525.1K/2.0M/613.5K	527.3K/1.8M/563.4K	441.5K/1.1M/510.0K	485.3K/2.4M/540.6K

Table 6: Number of (fully translated documents/ unique sentences/ unique words) per language.

5. Bibliographical References

Cortes, C. and Vapnik, V. (1995). Support vector machine. *Machine learning*, 20(3):273–297.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Olsson, J. S., Oard, D. W., and Hajič, J. (2005). Cross-language text classification. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 645–646. ACM.